

Artificial Intelligence-Based Identification of Crop Yield Variations Using Agricultural and Weather Data

Dr. Javier Martínez

Assistant Professor, Institute of Agricultural Forecasting, Spain

Abstract

Crop yield varies in response to complex interactions among weather conditions, soil properties, crop-management practices, and seasonal growing conditions. Conventional statistical approaches can identify important associations between selected variables and yield, but agricultural relationships are frequently nonlinear and may involve interactions that are difficult to represent through simple parametric equations. Artificial intelligence provides a methodological approach for combining heterogeneous agricultural and weather information and identifying patterns associated with high, moderate, or reduced crop productivity. This hypothetical methodological study examines artificial intelligence-based identification of crop-yield variation using agricultural and meteorological variables. A fictional dataset containing 3,200 field-season observations is divided into training, validation, and independent testing subsets. Predictor variables include cumulative rainfall, maximum and minimum temperature, growing-degree accumulation, solar radiation, soil-moisture index, soil organic matter, nitrogen input, irrigation frequency, planting-date deviation, and previous-season yield.

Multiple linear regression, artificial neural network, and random forest regression are compared. The illustrative analysis evaluates prediction through root mean squared error, mean absolute error, and coefficient of determination. Under the assigned hypothetical values, random forest achieves the strongest overall performance, followed by the artificial neural network. A second analysis evaluates the random-forest model across moisture-limited, near-normal, heat-stressed, and favorable agro-weather conditions. Prediction errors are larger under stress-related conditions, illustrating the greater uncertainty associated with extreme environmental combinations. The study concludes that artificial intelligence can support identification of crop-yield variations when agricultural-management variables and weather information are analyzed jointly. However, reliable prediction requires representative multi-season data, appropriate temporal and geographic validation, agronomically meaningful variables, careful treatment of extreme weather, and clear distinction between statistical association and causal crop response. All numerical findings in this paper are illustrative and require empirical validation with verified agricultural data.

Keywords: Artificial Intelligence, Crop Yield Prediction, Agricultural Data, Weather Data, Machine Learning, Random Forest, Artificial Neural Network, Precision Agriculture, Agroclimatic Analytics, Yield Variability.

1. Introduction

Crop yield is influenced by interacting environmental and management factors operating throughout the growing season. Temperature, rainfall, solar radiation, soil moisture, nutrient availability, planting time, irrigation, soil characteristics, and crop-management practices can each contribute to observed yield differences. The effect of one variable may also depend on another; for example, elevated temperature can produce different crop responses under adequate and limited soil moisture. Accurate identification of yield variation therefore requires analysis capable of integrating multiple information sources.

Weather is particularly important because crop growth is strongly influenced by seasonal temperature and water availability. Lobell and Field examined global crop-yield responses to climate variation, while Lobell, Schlenker, and Costa-Roberts later demonstrated measurable relationships between climate trends and global crop production. Such research establishes that agricultural production cannot be interpreted independently of meteorological conditions. However, yield responses can vary according to crop type, growth stage, soil characteristics, and management practices.

Statistical crop models have traditionally been used to estimate relationships between yield and environmental predictors. Lobell and Burke examined the strengths and limitations of statistical models for estimating crop-yield responses to changing temperature and precipitation. Process-based crop models provide another important approach by representing physiological crop development and soil–plant interactions. Systems such as APSIM allow simulation of farming systems under changing management and environmental conditions but require extensive parameterization and domain-specific calibration.

Artificial intelligence and machine learning provide complementary data-driven methods. Drummond et al. compared statistical and neural-network approaches for site-specific yield prediction and demonstrated the potential of nonlinear learning when relationships among soil, topography, climate, and yield are complex. Everingham et al. applied ensemble data-mining approaches to regional sugarcane forecasting. Pantazi et al. later examined machine-learning techniques for wheat yield prediction using sensor-derived information, while Jeong et al. demonstrated strong random-forest performance across several crop-yield prediction settings.

The present hypothetical study develops a comparative AI framework for identifying variations in crop yield from agricultural and weather information. Multiple linear regression provides a conventional statistical benchmark, while artificial neural network and random forest models represent nonlinear machine-learning approaches. The study evaluates overall predictive performance and model behavior under contrasting agro-weather conditions. No actual farm, crop field, weather station, or production record is represented in the numerical analysis.

2. Objectives of the Study

The principal objective of this study is to examine how artificial intelligence can identify crop-yield variations through combined agricultural and weather data. The investigation emphasizes multivariable yield prediction, nonlinear agricultural relationships, weather-sensitive performance, and methodological requirements for reliable decision-support applications.

2.1 Identification of Agricultural and Weather Indicators

The first objective is to examine agricultural and meteorological variables that may contribute to crop-yield variation. The conceptual predictor set includes cumulative growing-season rainfall, temperature

extremes, growing-degree accumulation, solar radiation, soil moisture, soil organic matter, nitrogen application, irrigation frequency, planting timing, and previous-season productivity. These variables represent environmental conditions, soil characteristics, and controllable management factors rather than relying exclusively on one information source.

Yield responses may depend on the timing and interaction of these indicators. Total seasonal rainfall, for example, can conceal periods of water deficit during sensitive development stages, while average temperature can conceal damaging heat extremes. The methodology therefore conceptualizes both aggregated seasonal variables and selected stress-related measures. AI models are expected to identify statistical interactions among these variables, but such associations should not automatically be interpreted as causal agronomic relationships.

2.2 Comparison of Artificial Intelligence Models

The second objective is to compare a conventional linear model with nonlinear artificial-intelligence approaches. Multiple linear regression provides an interpretable benchmark in which changes in predictors are represented through fixed coefficients. The artificial neural network models nonlinear combinations through hidden computational units, while random forest regression combines multiple decision trees capable of capturing interactions and threshold-type relationships among weather and management variables.

The three models are compared through identical chronological and geographic data partitions. RMSE, MAE, and R^2 are used to measure prediction error and explanatory performance. This comparison is intended to determine whether nonlinear modeling provides additional predictive value under the hypothetical dataset. Model complexity is not considered beneficial by itself; an AI model should outperform simpler alternatives on genuinely unseen data before additional computational complexity is justified.

2.3 Evaluation of Yield Variation Under Contrasting Conditions

The third objective is to examine whether prediction performance changes under different agro-weather environments. Yield forecasting may be relatively straightforward during seasons resembling the historical average but substantially more difficult under drought, heat stress, excess rainfall, or unusual combinations of environmental variables. Model evaluation should therefore distinguish average accuracy from robustness under stressful or less frequently represented conditions.

The study categorizes the hypothetical evaluation observations into moisture-limited, near-normal, heat-stressed, and favorable agro-weather conditions. These categories are used only for performance analysis and are not treated as universal agronomic definitions. The objective is to demonstrate that AI-based crop-yield models should be examined under multiple environmental regimes because high average performance can conceal weaker behavior during agriculturally important extremes.

3. Review of Literature

3.1 Weather, Climate, and Crop-Yield Variation

Lobell and Field investigated relationships between global-scale climate variation and crop yields, demonstrating that temperature and precipitation contribute meaningfully to agricultural production changes. Lobell et al. subsequently analyzed climate trends and global crop production and found

evidence that changes in temperature had already affected yields of several major crops. These studies established an important empirical basis for incorporating weather information into agricultural productivity analysis.

Schlenker and Roberts examined nonlinear relationships between temperature and crop yields and showed that extreme heat can produce disproportionately adverse responses. Lobell and Burke further evaluated statistical modeling approaches for predicting crop-yield responses to changing temperature and precipitation. Collectively, this research demonstrates that crop response cannot always be approximated adequately by simple linear averages and that weather extremes and nonlinear effects deserve explicit consideration.

3.2 Artificial Intelligence and Machine Learning for Yield Prediction

Drummond et al. compared multiple linear, projection-pursuit, and neural-network approaches for site-specific yield prediction using soil, terrain, climate, and yield information. Their results demonstrated the potential of neural methods to represent complex agricultural relationships, although performance depended on the amount and diversity of available information. This work provided an important early demonstration of machine learning within precision crop-yield modeling.

Everingham et al. used ensemble data-mining methods to forecast regional sugarcane production under heterogeneous environmental and management conditions. Pantazi et al. later evaluated computational-intelligence models for wheat yield prediction using high-resolution agricultural sensing data. Jeong et al. compared random forests with multiple linear regression for wheat, maize, and potato yield prediction and demonstrated the potential of ensemble tree methods across regional and global prediction problems.

3.3 Crop Simulation, Precision Agriculture, and Integrated Data

Keating et al. described APSIM as a flexible farming-systems simulation environment capable of representing interactions among crops, soils, climate, and agricultural management. Process-based simulation models provide mechanistic insight and can examine scenarios not represented directly in historical observations. However, they differ fundamentally from machine-learning approaches, which infer statistical relationships directly from data and can require less explicit physiological specification.

Precision-agriculture research has further demonstrated the value of combining spatial soil, crop, and management information. Kitchen et al. and related site-specific research examined relationships among soil properties, landscape position, and crop productivity. Such work shows that weather alone cannot fully explain within- and between-field yield differences. Integrated AI models therefore benefit from considering both environmental conditions and management or soil variables when these data are available and reliable.

4. Research Methodology

4.1 Research Design and Predictive Models

The present study adopts a hypothetical supervised regression design. Three predictive models are considered: multiple linear regression, a feed-forward artificial neural network, and random forest regression. The linear model serves as a conventional benchmark. The neural network conceptually contains two hidden layers with nonlinear activation functions, while random forest combines multiple regression trees trained on bootstrapped samples and randomly selected feature subsets.

Model development is performed using identical records and predictor definitions. Hyperparameters are assumed to be selected using the validation subset only. The independent test data remain unavailable during model fitting and model selection. This design reduces optimistic bias and permits a direct comparison among model families. No actual software implementation or machine-learning training was performed to generate the results presented in the paper.

4.2 Hypothetical Agricultural and Weather Dataset

The fictional dataset contains 3,200 field-season records representing multiple growing seasons and agricultural production zones. Two thousand two hundred forty observations are assigned to model training, 480 to validation, and 480 to independent testing. Yield is represented hypothetically in metric tonnes per hectare. The records do not correspond to any real farmer, agricultural company, research station, weather service, or geographic location.

Predictor variables include growing-season rainfall, maximum-temperature frequency, minimum temperature, growing-degree accumulation, solar radiation, soil-moisture index, soil organic matter, soil pH, nitrogen input, irrigation frequency, planting-date deviation, previous-season yield, and selected interaction-ready management indicators. Weather variables are assumed to be temporally aligned with the corresponding crop season. Records containing future harvest information are excluded from predictor construction to prevent information leakage.

4.3 Evaluation Metrics and Analytical Procedure

Predictive performance is evaluated using RMSE, MAE, and the coefficient of determination. RMSE is calculated as $RMSE = \sqrt{[(1/n)\sum(y_i - \hat{y}_i)^2]}$, while MAE is $MAE = (1/n)\sum|y_i - \hat{y}_i|$. Lower RMSE and MAE indicate smaller prediction errors, while a larger R^2 indicates that a greater proportion of observed yield variation is represented by model predictions.

A second analysis evaluates the strongest hypothetical model across four agro-weather categories constructed from moisture and temperature indicators. Model error is examined separately under each condition. In real agricultural research, repeated spatial and temporal cross-validation, crop-specific calibration, uncertainty intervals, and independent multi-year evaluation would also be necessary. The present numerical values are constructed solely to demonstrate appropriate reporting and interpretation.

5. Analysis

5.1 Comparative Crop-Yield Prediction Performance

The first hypothetical analysis compares overall predictive performance across the 480-observation test set. Multiple linear regression records an RMSE of 0.78 t/ha and R^2 of 0.68, indicating that a linear combination of agricultural and weather variables captures substantial but incomplete yield variation. The artificial neural network improves the assigned RMSE to 0.59 t/ha, while random forest produces the lowest illustrative RMSE at 0.53 t/ha.

The random forest also records the lowest MAE and the highest R^2 . This pattern conceptually reflects the ability of tree ensembles to represent nonlinear thresholds and interactions among temperature, moisture, soil, and management variables. However, the assigned differences are entirely hypothetical. Actual relative performance would depend on the crop, geographic scale, environmental variability, sample size, predictor quality, and model tuning procedure.

Table 1: Hypothetical Comparative Crop-Yield Prediction Performance

Predictive Model	RMSE (t/ha)	MAE (t/ha)	R ²
Multiple Linear Regression	0.78	0.61	0.68
Artificial Neural Network	0.59	0.45	0.82
Random Forest Regression	0.53	0.40	0.86

5.2 Prediction Under Different Agro-Weather Conditions

The second hypothetical analysis evaluates random-forest performance under four environmental conditions. Near-normal growing conditions produce the smallest assigned prediction error, with RMSE of 0.44 t/ha and R² of 0.89. Favorable moisture and temperature conditions also produce comparatively strong prediction. Moisture-limited and heat-stressed observations show higher errors and lower R² values, suggesting that extreme or less frequently represented conditions are conceptually more difficult to model.

The pattern illustrates why overall test accuracy can provide an incomplete account of agricultural model reliability. Yield losses during drought or extreme heat may be operationally important even when such observations represent a minority of the dataset. Prediction models should therefore be evaluated explicitly on stressful seasons and unusual environmental combinations. A model that performs strongly in average years but poorly during crop-stress conditions may provide limited value for risk-oriented agricultural planning.

Table 2: Hypothetical Random-Forest Performance Across Agro-Weather Conditions

Agro-Weather Condition	RMSE (t/ha)	MAE (t/ha)	R ²
Moisture-Limited Conditions	0.62	0.48	0.80
Near-Normal Conditions	0.44	0.34	0.89
Heat-Stressed Conditions	0.66	0.51	0.78
Favorable Moisture–Temperature Conditions	0.47	0.36	0.88

6. Discussion

6.1 Nonlinear Agricultural Relationships and AI

The hypothetical findings illustrate why nonlinear machine-learning models may provide useful advantages in crop-yield prediction. Agricultural production is rarely controlled by independent additive effects alone. Temperature response can depend on soil moisture, irrigation can modify drought sensitivity, and nutrient availability can interact with rainfall or soil characteristics. Artificial neural networks and random forests can represent such interactions without requiring every nonlinear relationship to be defined manually.

This interpretation is consistent with Drummond et al., Pantazi et al., and Jeong et al., who demonstrated the value of machine-learning or neural methods within agricultural yield analysis. Nevertheless, improved prediction does not prove that a model has identified the biological mechanism

governing crop response. Machine learning can exploit correlations that may change across locations and years. Agronomic knowledge remains necessary when interpreting influential variables and transferring models to new production environments.

6.2 Weather Information and Yield Variability

The regime-specific analysis emphasizes the importance of weather variation in crop-yield modeling. Temperature and rainfall influence crop development throughout the growing season, while extremes can produce effects that differ substantially from average conditions. Lobell and Field and Schlenker and Roberts demonstrated that crop yields can respond nonlinearly to temperature, reinforcing the need to represent stress-related weather information rather than relying only on seasonal averages.

Weather effects must also be considered together with management. Irrigation can reduce the yield consequences of rainfall shortage, while planting date can alter exposure to high temperatures during sensitive developmental stages. Consequently, an AI model that uses weather but ignores agricultural practices may attribute too much variation to environmental variables. Integrated datasets provide a stronger methodological basis because they allow the model to represent both external conditions and management responses.

6.3 Prediction Performance and Agricultural Interpretation

RMSE and MAE provide useful information concerning average numerical prediction error, while R^2 measures the proportion of observed variation represented by the predictions. None of these metrics alone determines whether a model is useful for agricultural decision-making. An error of 0.5 t/ha may be acceptable for one crop or planning task but excessive for another. Evaluation should therefore be connected to the intended spatial scale and agricultural decision.

Yield prediction should also be distinguished from causal assessment. A random forest may identify rainfall or nitrogen input as highly predictive without establishing what yield would have occurred if those variables had been changed independently. Causal agronomic recommendations generally require experimental design, mechanistic understanding, or suitable causal-inference methods. Machine-learning yield prediction is most defensible when used for forecasting and pattern identification rather than unqualified prescriptions about agricultural inputs.

7. Challenges

7.1 Weather Extremes and Limited Historical Representation

Extreme drought, heat, excessive rainfall, and unusual seasonal sequences occur less frequently than ordinary conditions and may therefore be underrepresented in training datasets. Machine-learning models generally perform most reliably within the range of conditions represented during development. Jeong et al. noted that random-forest prediction can lose accuracy near extreme ends or beyond the boundaries of training information. This limitation is particularly important for agriculture because extreme conditions often produce the largest yield losses.

Researchers should therefore examine the environmental coverage of training datasets before applying models to unusual seasons. Temporal cross-validation can reveal whether prediction quality remains stable across contrasting years. Combining observational data with process-based simulation may also provide a useful research direction where historical examples of extreme conditions are scarce, although simulated observations must be treated differently from measured agricultural outcomes.

7.2 Geographic Transfer and Management Heterogeneity

Agricultural relationships differ among soil types, crop cultivars, management systems, irrigation practices, and geographic regions. A model developed from one production zone may therefore perform poorly elsewhere even when the same crop is grown. Random field-level splitting can exaggerate generalization if neighboring observations share soil, climate, and management characteristics. Geographic separation is consequently essential for evaluating transferable yield prediction.

Management heterogeneity creates additional complexity because available variables may not fully represent farmer decisions. Irrigation timing, fertilizer formulation, pest pressure, planting density, and cultivar selection can all affect yield. Missing management information may cause the model to attribute unexplained variation incorrectly to weather or soil variables. High-quality AI prediction therefore depends as much on comprehensive agricultural records as on algorithm selection.

7.3 Data Quality, Timing, and Interpretability

Weather stations, field sensors, farm-management records, and yield-monitoring systems can contain missing observations, calibration errors, inconsistent units, and temporal misalignment. A rainfall measurement from the wrong spatial location or a sensor reading outside the relevant growth stage can introduce substantial noise. Data-cleaning procedures should therefore be documented transparently and performed without using future outcome information.

Interpretability also remains important when predictions are used in agricultural decision support. Tree-based variable importance, partial-dependence analysis, or sensitivity analysis can help identify which features influence predictions, but these techniques must be interpreted cautiously when predictors are correlated. A useful AI system should provide agronomically understandable information while acknowledging uncertainty rather than presenting yield estimates as exact forecasts.

8. Conclusion

Artificial intelligence provides a useful methodological framework for identifying crop-yield variations from integrated agricultural and weather information. The present hypothetical study compared multiple linear regression, artificial neural network, and random forest regression using environmental, soil, and management indicators. The design emphasizes that crop productivity should be modeled as the outcome of interacting agroclimatic and agricultural conditions rather than as a response to one isolated variable.

The illustrative results assigned the strongest overall prediction performance to random forest, with an RMSE of 0.53 t/ha and R^2 of 0.86. The model produced larger hypothetical errors under moisture-limited and heat-stressed conditions than under near-normal or favorable conditions. These results are entirely constructed and provide no empirical evidence concerning any actual crop, farm, region, or growing season.

The study concludes that future empirical crop-yield prediction should use geographically and temporally diverse observations, integrate management and weather information, evaluate performance under extreme conditions, and compare AI methods with appropriate statistical or process-based baselines. Machine learning can support yield forecasting and variation identification, but reliable agricultural application requires agronomic validation, transparent uncertainty, and careful testing beyond the conditions used for model development.

References

1. Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Computers and Electronics in Agriculture*, 155, 257–282. <https://doi.org/10.1016/j.compag.2018.10.024>
2. Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, Article 621. <https://doi.org/10.3389/fpls.2019.00621>
3. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, Article 105709. <https://doi.org/10.1016/j.compag.2020.105709>
4. Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14(9), Article 1990. <https://doi.org/10.3390/rs14091990>
5. Javed, M. A., & Murad, M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, 10(24), Article e40836. <https://doi.org/10.1016/j.heliyon.2024.e40836>
6. Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11), 1443–1452. <https://doi.org/10.1016/j.agrformet.2010.07.008>
7. Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Agricultural and Forest Meteorology*, 250–251, 242–251. <https://doi.org/10.1016/j.agrformet.2018.03.024>
8. Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random forests for global and regional crop yield predictions. *PLOS ONE*, 11(6), Article e0156571. <https://doi.org/10.1371/journal.pone.0156571>
9. You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). <https://doi.org/10.1609/aaai.v31i1.11172>
10. Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, 10, Article 1750. <https://doi.org/10.3389/fpls.2019.01750>
11. Gopal, P. M., & Bhargavi, R. (2019). Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. *Applied Artificial Intelligence*, 33(7), 621–642. <https://doi.org/10.1080/08839514.2019.1592343>
12. Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019). Crop yield prediction using machine learning algorithms. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*. IEEE. <https://doi.org/10.1109/ICIIP47207.2019.8985951>
13. Elavarasan, D., & Vincent, P. M. D. R. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access*, 8, 86886–86901. <https://doi.org/10.1109/ACCESS.2020.2992480>
14. Cheng, E., Zhang, B., Peng, D., et al. (2022). Wheat yield estimation using remote sensing data based on machine learning approaches. *Frontiers in Plant Science*, 13, Article 1090970. <https://doi.org/10.3389/fpls.2022.1090970>

15. Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14(9), Article 1990. <https://doi.org/10.3390/rs14091990>
16. Patel, N. R., Pokhariyal, S., & Singh, R. P. (2023). Advancements in remote sensing based crop yield modelling in India. *Journal of Agrometeorology*, 25(3). <https://doi.org/10.54386/jam.v25i3.2316>
17. Crop yield prediction using machine learning approaches on a wide spectrum. (2022). *Computers, Materials & Continua*, 72(3), 5663–5679. <https://doi.org/10.32604/cmc.2022.027178>
18. Crop yield prediction using machine learning techniques. (2023). *Advanced Engineering Informatics*, 55, Article 103326.
19. Crop yield prediction using machine learning: An extensive and systematic literature review. (2025). *Smart Agricultural Technology*, 10, Article 100718. <https://doi.org/10.1016/j.atech.2024.100718>
20. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. (2019). *Agricultural and Forest Meteorology*, 264, 1–15. <https://doi.org/10.1016/j.agrformet.2018.09.021>